

ALE Analytics - A microbial data warehouse and analytical suite

Patrick Phaneuf¹, Dennis Gosting², Bernhard O. Palsson^{2,3}, Adam Feist^{2,3}

1) Department of Computer Science and Engineering, University of California at San Diego, La Jolla 92093-0404

2) Department of Bioengineering, University of California San Diego, La Jolla, CA, 92093, United States of America,

3) Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800, Lyngby, Denmark

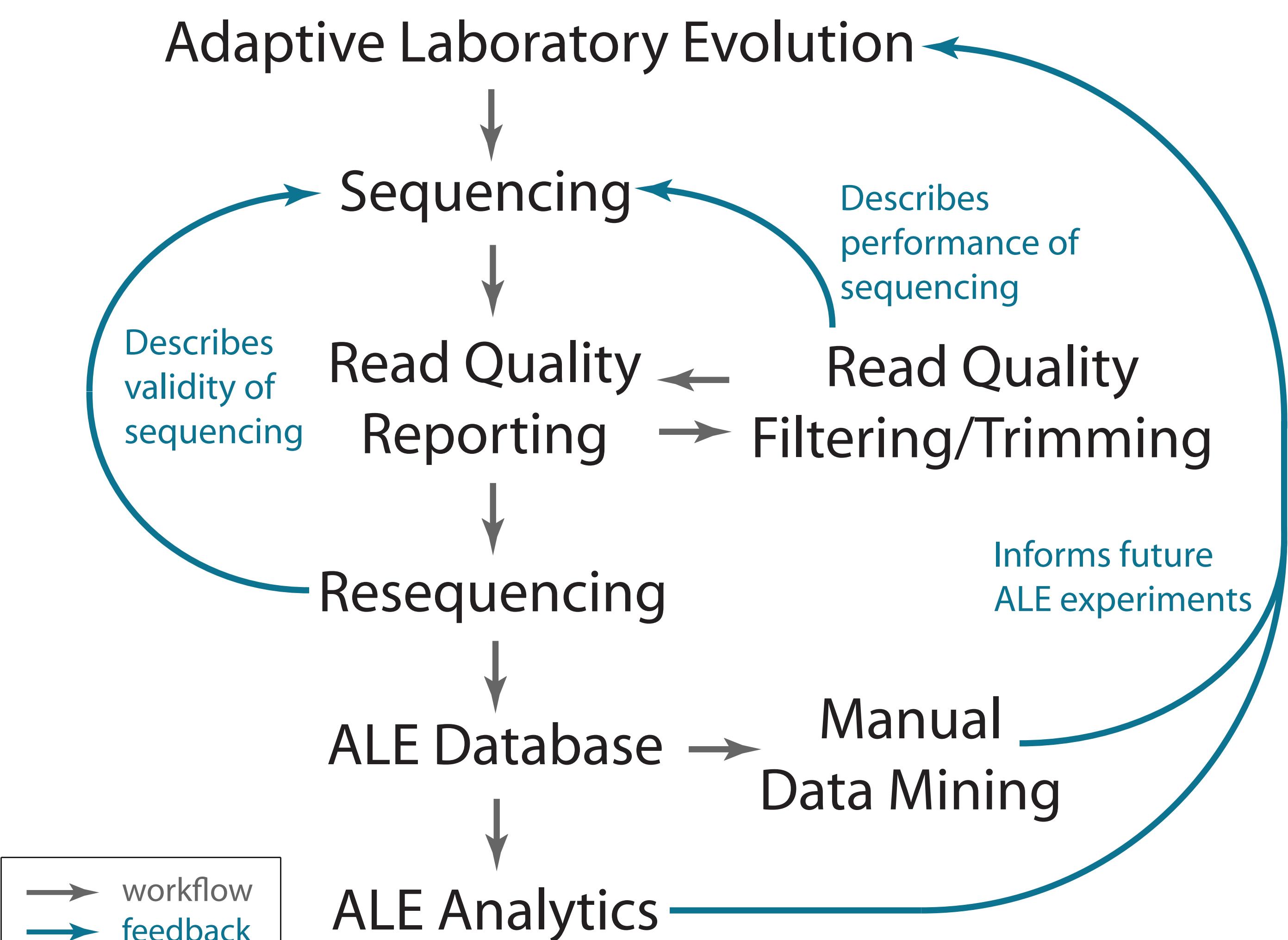
Abstract

Adaptive Laboratory Evolution (ALE) is a powerful experimental tool used by academic and industrial labs for a number of applications. ALE experiments generate a substantial amount of data, coming in the form of sequencing reads, alignment reports, and sample metadata. As ALE experiments scale to include more samples, the task of managing this data comes at higher costs due to the effort necessary to organize and integrate data into a format that describes the evolution process succinctly. In this work, we describe the development, deployment and iteration of an 'ALE Analytics' pipeline and web platform that streamlines the necessary ALE experiment data post-processing, manages experiment data, and produces interactive reports that detail an ALE experiment. Our design has been primarily driven by the need to consolidate large amounts of ALE experimental data in such a way to describe the quality of the sample sequencing, adaptive mutations in evolved strains, the context of mutations via their metadata (i.e., culturing environments, strain properties), and related mutations found in other experiments housed in the database. We have done so by leveraging a full stack of technologies that enable the parsing and databasing of experiment data, the execution of automated analysis on said data and the generation of web accessible reports. Future efforts will take full advantage of this developing platform to enable more depth and breadth of ALE experiment analysis with quicker turnaround.

ALE Mutation Processing Pipeline

Description of ALE Analytics goal

Bacon ipsum dolor amet ground round ribeye pastrami pork chop, pork belly fatback kevin. Pork picanha kielbasa short ribs, biltong sausage kevin pastrami sirloin porchetta pork chop filet mignon



Features and Advantages

- Multiple bacterial strain support.
- Industry established workflow with industry established tools.
- Automated multi-sample processing for specific stages.
- Stage dependent feedback.
- Feedback provides opportunity for experiment and data refinement at specific stages.

ALE Analytics Web Platform

Description of ALE Analytics goal

Bacon ipsum dolor amet ground round ribeye pastrami pork chop, pork belly fatback kevin. Pork picanha kielbasa short ribs, biltong sausage kevin pastrami sirloin porchetta pork chop filet mignon

Whole database trend reporting

Dashboard

General statistics on entire ALE Database including:

- Most frequently mutated genes
- Most frequent mutations
- Mutation type counts (SNP, INDEL, etc.)

Genes

Gene specific information:

- All mutations in the ALE Database involving a particular gene.
- 3D rendering of a gene's product.
- Highlight structural area of a gene's product effected by a mutation.
- Display's ligand structure at site of mutation

Mutation Search

Search the entire ALE Database by mutation specific features.

Contextualizing mutations

Sample Metadata

A description of each ALE sample:

- strain.
- genetic perturbations,
- media,
- substrate,
- temperature,
- etc

Shared Key Mutations and Metadata

Key mutations shared among multiple ALE experiments and their experimental conditions.

ALE Analytics

A user web interface reporting ALE experiment data and trends found through automated datamining.

ALE Database

Database containing the mutational data and metadata generated by the ALE data processing pipeline. This database is the basis for all analytical features implemented within the ALE Analytics web platform

Experiment Statistics

Statistics and visualizations that can lend intuition on an ALE experiment:

- Sequencing data quality.
- Mutational "hotspots".
- Most frequently mutated genes.
- Most frequent mutations.

Mutation trend reporting

Fixating Mutations

A report of ALE samples and their mutations that emerged and fixated within an ALE. Presents which mutations were kept by the dominant populations between flasks and therefore may have lead to fitness benefits.

Frequently Mutated Genes

A report of an ALE experiment's samples and their sets of mutations that:

1. Affect the same gene within a sample
 2. Affect the same gene among different samples
- Highlights which genes are more frequently mutated within an ALE experiment, potentially elucidating the functional categories that were more perturbed.

Common Mutations

A report of ALE samples and their sets of shared mutations. Used to find samples most mutationally similar to a currently selected sample.

Mutation Lineages

A report of ALE samples and their mutations. Fundamental feature used in understanding an ALE's mutational evolution, where one visually inspects for mutational trends, such as fixation, among all samples within an ALE experiment.

Duplications

Genomic duplication and amplification finding based on resequencing alignment read-depth. Can find large areas of duplication that are not otherwise found with current resequencing tools.

References

Original paper and those I report biological findings on.
Additionally, any 3rd party software I mention.
All papers that I've taken figures from.

Future Directions

Integrate ALE experiment fitness data to provide additional context for identifying causal mutations.

jump regions
averaged data
smoothed data

mutations that appear and persist in all following flasks
mutations that appear and do not persist

